Choix de modèle pour les champs de Gibbs par un algorithme ABC.

Application à la prédiction de la structure 3D d'une protéine.

Aude Grelaud[1,2,3] & Christian P. Robert[2,3] & Jean-Michel Marin[4,3] & François Rodolphe[1] & Jean-François Taly[1]

[1] *INRA, unité MIG, Domaine du Vilvert, 78350 Jouy-en-Josas, France.*
[2] *CEREMADE, Université Paris Dauphine, Place du Maréchal De Lattre De Tassigny, 75775 Paris cedex 16, France.*
[3] *CREST-INSEE, Timbre J340, 3, Avenue Pierre Larousse, 92240 Malakoff, France.*
[4] *I3M, Université Montpellier 2, Case courrier 051, Place Eugène Bataillon, 34095 Montpellier cedex, France.*

Les champs de Gibbs sont des modèles souvent utilisés pour l'analyse de données présentant des corrélations spatiales, notamment en analyse d'image [4]. La définition du modèle est alors liée à une système de voisinage. Nous considérons ici des champs de Gibbs associés à une fonction de densité de la forme suivante

$$f(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z_{\boldsymbol{\theta}}} \exp\{\boldsymbol{\theta}^T S(\mathbf{x})\}\,,$$

où $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$ ($\mathcal{X}$ étant un ensemble fini), $\boldsymbol{\theta} \in \mathbb{R}^p$ est le paramètre associé au modèle et $S(.)$ est une fonction à valeurs dans $\mathbb{R}^p$. $\boldsymbol{\theta}^T S(\cdot)$ est le potentiel définissant le champ de Gibbs, la fonction $S(.)$ étant étroitement liée à la structure de voisinage au sens où $S(\mathbf{x}) = \sum_{i\sim i'} S^{i,j}(x_i, x_{i'})$, $i \sim i'$ signifiant que i et $i^{'}$ sont voisins. De plus, $S(\mathbf{x})$ constitue une statistique exhaustive pour le paramètre $\boldsymbol{\theta}$. Enfin, $Z_{\boldsymbol{\theta}} > 0$ est la constante de normalisation, impossible à calculer dans la majorité des cas, ce qui complique l'inférence sur le paramètre.
Notre objectif est de choisir au travers d'un échantillon le modèle le plus approprié parmi un ensemble de M champs de Gibbs de densités

$$f_m(\mathbf{x}|\boldsymbol{\theta}_m) = \exp\left\{\boldsymbol{\theta}_m^T S_m(\mathbf{x})\right\} / Z_{\boldsymbol{\theta}_m,m}\,,$$

où $S_m(.)$ et $Z_{\boldsymbol{\theta}_m,m}$ sont respectivement la statistique exhaustive et la constante de normalisation correspondant au modèle m, $1 \leq m \leq M$. Encore une fois, on peut réécrire $S_m(\mathbf{x})$ sous la forme $\sum_{i\overset{m}{\sim}i'} S^{i,j}(x_i, x_{i'})$. Le problème peut être vu comme un choix entre M fonctions S_m, $1 \leq m \leq M$ ou entre M structures de voisinage.
Nous considérons un nouveau vecteur de paramètres $(\mathcal{M}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ qui inclut l'indice du modèle $\mathcal{M}$. Dans un cadre bayésien, nous définissons une distribution *a priori* pour l'indice du modèle, $\pi(\mathcal{M} = m)$, ainsi que pour le paramètre conditionnellement à la valeur

m de l'indice du modèle, $\pi_m(\boldsymbol{\theta}_m)$, définie sur l'espace Θ_m. Le choix du modèle repose alors sur les probabilités *a posteriori* des modèles

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) \propto \int_{\Theta_m} f_m(\mathbf{x}|\boldsymbol{\theta}_m)\pi_m(\boldsymbol{\theta}_m)\,\mathrm{d}\boldsymbol{\theta}_m\,\pi(\mathcal{M} = m)\ .$$

De celles-ci, on peut tirer le facteur de Bayes. Les méthodes utilisées habituellement pour estimer ces quantités nécessitent de calculer exactement la vraisemblance (voir [2] par exemple) et ne sont pas applicables ici car la constante de normalisation n'est pas disponible.
Récemment, des méthodes permettant de faire de l'inférence sur les paramètres sans utiliser la vraisemblance sont apparues (voir [7], [1] et [6]). L'algorithme ABC (Approximate Bayesian Computation) [1] est le suivant : pour un vecteur de données $\mathbf{x}^0$ de distribution $f(\mathbf{x}|\boldsymbol{\theta})$ et une distribution *a priori* $\pi(\boldsymbol{\theta})$ pour le paramètre $\boldsymbol{\theta}$, une valeur $\boldsymbol{\theta}^*$ est générée selon la distribution *a priori*, $\boldsymbol{\theta}^* \sim \pi(\cdot)$, puis une valeur $\mathbf{x}^*$ est générée selon $\mathbf{x}^* \sim f(\cdot|\boldsymbol{\theta}^*)$. On accepte les paramètres $\boldsymbol{\theta}^*$ pour lesquels $\rho(T(\mathbf{x}^0), T(\mathbf{x}^*)) < \epsilon$ où $T(\cdot)$ est une statistique résumée, $\rho(\cdot,\cdot)$ une distance et ϵ une tolérance. L'algorithme ABC génère des paramètres de distribution $\pi\{\boldsymbol{\theta}|\rho(T(\mathbf{x}^*), T(\mathbf{x}^0)) < \epsilon\}$ qui est une bonne approximation de la distribution $\pi(\boldsymbol{\theta}|\mathbf{x}^0)$ lorsque ϵ est petit et $T(\cdot)$ est une statistique suffisamment représentative des données. L'idéal est de prendre pour $T(\mathbf{x})$ une statistique exhaustive pour le paramètre $\boldsymbol{\theta}$, mais, en pratique, ce choix est rarement disponible. Si $T(\mathbf{x})$ est une statistique exhaustive et $\epsilon = 0$, les paramètres $\boldsymbol{\theta}^*$ retenus ont exactement pour distribution $\pi(\boldsymbol{\theta}|\mathbf{x}^0)$.
Rappelons que l'objectif est d'estimer les probabilités *a posteriori* des modèles $\mathbb{P}(\mathcal{M} = m|\mathbf{x})$. L'algorithme ABC-MC génère un échantillon $(\boldsymbol{\theta}^{i*}_{m^{i*}}, m^{i*})_{(1\leq i\leq N)}$ dont la distribution est approximativement $\pi(m, \boldsymbol{\theta}_1, .., \boldsymbol{\theta}_M|\mathbf{x}^0)$. Une estimation de $\mathbb{P}(\mathcal{M} = m|\mathbf{x})$ fondée sur cet échantillon est donnée par la fréquence empirique des visites du modèle

$$\widehat{\mathbb{P}}(\mathcal{M} = m|\mathbf{x}^0) = \sharp\{m^{i*} = m\}/N\ ,\ 1 \leq m \leq M,$$

que l'on peut ensuite utiliser pour estimer le facteur de Bayes.
Nous appliquons cette procédure à une collection de champs de Gibbs se distinguant par leur système de voisinage. Nous choisissons pour vecteur de statistiques résumées $T(\mathbf{x})$ la concaténation des statistiques exhaustives de chacun des modèles $T(\mathbf{x}) = (S_1(\mathbf{x}), \cdots, S_M(\mathbf{x}))$ et montrons que c'est une statistique exhaustive pour le paramètre $(\mathcal{M}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$.

Nous avons étudié les performances de cette méthode sur deux cas particuliers de champs de Gibbs pour lesquels la constante de normalisation est disponible, dans le cas exact, puis en introduisant une tolérance ϵ. Dans les deux cas, nous obtenons une bonne approximation des probabilités *a posteriori* puis des facteurs de Bayes.

Nous avons ensuite utilisé la méthode ABC-MC pour prédire la structure 3D d'une protéine. La connaissance de la structure tridimensionnelle d'une protéine apporte de nombreuses informations sur celle-ci, notamment sur sa fonction. Ces dernières années,

les méthodes basées sur le repliement de la séquence de la protéine d'intérêt sur des structures connues, appelées *threading*, se sont énormément developpées [5]. Elles forment une alternative à la détermination expérimentale de la structure 3D moins coûteuse, moins délicate et surtout plus rapide. Elles fournissent non pas la structure réelle, mais un ensemble de structures candidates. L'inconvénient est que les critères de choix entre les propositions obtenues sont souvent insuffisants. L'idée ici est d'utiliser le fait que les acides aminés voisins dans la structure 3D ont souvent des propriétés biochimiques similaires pour construire un critère de choix. La propriété que nous utilisons ici est l'hydrophobicité. Ainsi, à chaque acide aminé est attribuée une valeur 0 ou 1 suivant qu'il soit hydrophile ou hydrophobe. De plus, chaque structure proposée définit une système de voisinage à partir duquel nous construisons un modèle de champ de Gibbs. Nous pouvons alors utiliser l'algorithme ABC-MC pour faire un choix. Nous appliquons cette procédure à une protéine de la bactérie *Thermotaga maritima*.

Mots clés : ABC, choix de modèle bayésien, champs de Gibbs, structure 3D de protéine.

Gibbs random fields are statistical models often used to analyse spatially correlated data, especially in image analysis [4]. Each model is associated with a dependence structure and has a likelihood function

$$f(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z_{\boldsymbol{\theta}}} \exp\{\boldsymbol{\theta}^{\mathrm{T}} S(\mathbf{x})\}\,,$$

where $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$ ($\mathcal{X}$ is finite), $\boldsymbol{\theta} \in \mathbb{R}^p$ is the parameter and $S(.)$ is a fonction taking values in $\mathbb{R}^p$. $\boldsymbol{\theta}^T S(\cdot)$ is the potential defining the Gibbs random field, $S(.)$ depending on the neighborhood structure in that $S(\mathbf{x}) = \sum_{i\sim i'} S^{i,j}(x_i, x_{i'})$, $i \sim i'$ meaning that i and $i^{'}$ are neighbours. Notice that $S(\mathbf{x})$ is a sufficient statistic for $\boldsymbol{\theta}$. $Z_{\boldsymbol{\theta}} > 0$ is the corresponding normalising constant and its unavailability complicates the inference on the parameter $\boldsymbol{\theta}$. Our target is to select the most appropriate model among M Gibbs random fields with likelihood function

$$f_m(\mathbf{x}|\boldsymbol{\theta}_m) = \exp\left\{\boldsymbol{\theta}_m^T S_m(\mathbf{x})\right\} / Z_{\boldsymbol{\theta}_m,m}\,,$$

where $S_m(.)$ and $Z_{\boldsymbol{\theta}_m,m}$ are respectively a sufficient statistic for the parameter $\boldsymbol{\theta}_m$ and the normalising constant corresponding to model m, $1 \leq m \leq M$. Once more, $S_m(\mathbf{x})$ can be written as $\sum_{i \overset{m}{\sim} i'} S^{i,j}(x_i, x_{i'})$. The choice is thus between M functions S_m, $1 \leq m \leq M$ or M neighborhood structures.
We consider a new parameter $(\mathcal{M}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ including the model index $\mathcal{M}$. In a Bayesian framework, we define a prior distribution on the model index, $\pi(\mathcal{M} = m)$, and on the parameter given the model index m, $\pi_m(\boldsymbol{\theta}_m)$. The model choice relies on the posterior probabilities of each model

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) \propto \int_{\Theta_m} f_m(\mathbf{x}|\boldsymbol{\theta}_m)\pi_m(\boldsymbol{\theta}_m)\,\mathrm{d}\boldsymbol{\theta}_m\,\pi(\mathcal{M} = m)\ .$$

these can be used to calculate Bayes factors. Standard methods of estimation cannot apply since the likelihood is available only up to a normalising constant [2].
Recently, new methods have been developed to evaluate posterior distributions when the likelihood function is analytically or computationaly intractable (introduced by [7] and extended in [1] and [6]). The ABC (Approximate Bayesian Computation) algorithm [1] is the following: given a dataset $\mathbf{x}^0$ associated with the sampling distribution $f(\cdot|\boldsymbol{\theta})$, and under a prior distribution $\pi(\boldsymbol{\theta})$ on the parameter $\boldsymbol{\theta}$, jointly generate a parameter value $\boldsymbol{\theta}^*$ from the prior, $\boldsymbol{\theta}^* \sim \pi(\cdot)$ and a value $\mathbf{x}^*$ from the sampling distribution $\mathbf{x}^* \sim f(\cdot|\boldsymbol{\theta}^*)$. $\boldsymbol{\theta}^*$ is accepted when $\rho(T(\mathbf{x}^0), T(\mathbf{x}^*)) < \epsilon$ with $T(\cdot)$ a summary statistic, $\rho(\cdot,\cdot)$ a distance and ϵ a fixed tolerance. This algorithm samples from the distribution $\pi\{\boldsymbol{\theta}|\rho(T(\mathbf{x}^*), T(\mathbf{x}^0)) < \epsilon\}$ which is a good approximation of the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x}^0)$ when ϵ is small et $T(\cdot)$ is close to being sufficient. The best choice is to pick $T(\mathbf{x})$ as a sufficient statistic, but this choice is rarely available. When $T(\mathbf{x})$ is a sufficient statistic and $\epsilon = 0$, the ABC algorithm samples exactly from the distribution $\pi(\boldsymbol{\theta}|\mathbf{x}^0)$.
Our aim is to evaluate the posterior model probabilities $\mathbb{P}(\mathcal{M} = m|\mathbf{x})$. The ABC-MC algorithm (MC stands for Model Choice) generate a sample $(\boldsymbol{\theta}^{i*}_{m^{i*}}, m^{i*})_{(1\leq i\leq N)}$ which distribution is approximatively $\pi(m, \boldsymbol{\theta}_1, .., \boldsymbol{\theta}_M|\mathbf{x}^0)$. An approximation of the posterior probabilities based on a sample of N values $(\boldsymbol{\theta}^{i*}_{m^{i*}}, m^{i*})$, $(1 \leq i \leq N)$, generated from this algorithm is the empirical frequencies of visits to the models, namely

$$\widehat{\mathbb{P}}(\mathcal{M} = m|\mathbf{x}^0) = \sharp\{m^{i*} = m\}/N \ , \ 1 \leq m \leq M,$$

which can be plugged-in to estimate a Bayes factor.
We apply this algorithm in the context of Gibbs random fields. The vector of summary statistics $T(\mathbf{x})$ is choosen as the concatenation of the sufficient statistic of each model $T(\mathbf{x}) = (S_1(\mathbf{x}), \cdots, S_M(\mathbf{x}))$.We show that $T(\mathbf{x})$ is a sufficient statistic for the joint parameter $(\mathcal{M}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$.

We describe the performances of this approach using two models which are special cases of Gibbs random fields with available normalising constants, in the exact case and with a tolerance ϵ. In both cases, we obtain good approximations.
Then, the ABC-MC approach was used to predict the 3D structure of a protein. The knowledge of a protein tridimensional structure provides important information, especially about its function. Recently, methods consisting in folding the sequence of a protein onto some known structures, called threading, have been developed [5]. These are an interessant alternative to experimental methods, less expensive and faster. The result is a list of candidate structures but the criterion given by this method usually cannot be used to make a choice between the propositions. We use here the fact that amino acids that are in contact in the 3D structure often have similar biochimical properties. In this application, a label (0 or 1) is allocated to each site (each amino acid) given its degree of hydrophobicity. In addition, each proposed structure is associated to a neighborhood structure and, thus, define a model. We are therefore in position to use the

ABC-MC approach to select a 3D structure among the candidate structures. We apply this procedure to a protein of the bacteria *Thermotaga maritima*.

Key words: ABC, Bayesian model choice, Gibbs random fields, protein 3D structure.

Bibliographie

[1] Beaumont, M., Zhang, W. et Balding, D. (2002) Approximate Bayesian Computation in population genetics, *Genetics*, 162, 2025–2035.
[2] Carlin, B.P. et Chib, S. (1995) Bayesian model choice via Markov chain Monte Carlo methods. *J. Royal Statist. Society Series B.*, 57, 473–484.
[3] Grelaud, A., Robert, C.P., Marin, J.-M., Rodolphe, F. et Taly, J.-F. (2008) ABC methods for model choice in Gibbs random fields. *arXiv:0807.2767*.
[4] Ibanez, M. and Simo, A. (2003) Parametric estimation in Markov random fields image modeling with imperfect observations. A comparative study. *Pattern Recognition Letters*, 24, 2377–2389.
[5] Marin, A., Pothier, J., Zimmermann, K., et Gibrat, J.F. (2002) FROST: a filterbased fold recognition method. *Proteins*, 49, 493–509.
[6] Marjoram, P., Molitor, J. Plagnol, V. et Tavaré, S. (2003) Markov chain Monte Carlo without likelihoods, *Proc. National Acad. Sci. USA.*, 100(26), 15324–15328.
[7] Pritchard, J. K., Seielstad, M. T., Perez-Lezaunand, A. et Feldman, M. W. (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.*, 16, 1791–1798.